Adaptive regularization framework for robust voice activity detection

Xugang Lu 1, Masashi Unoki 2, Ryosuke Isotani 1, Hisashi Kawai 1, Satoshi Nakamura 1

1National Institute of Information and Communications Technology, Japan
2Japan Advanced Institute of Science and Technology, Japan

Abstract

Traditional VAD algorithms work well under clean conditions, their performance however decreases drastically in noisy environments. We have investigated the tradeoff between false acceptance rate (FAR) and false rejection rate (FRR) in VAD with the consideration of noise reduction and speech distortion problem in speech enhancement, and proposed a regularization framework for noise reduction in designing VAD algorithms. In the framework, the balance between FAR and FRR was implicitly controlled by using a regularization parameter. In addition, the regularization was done in a reproducing kernel Hilbert space (RKHS) which made it easy to apply a nonlinear transform function via “kernel trick” for noise reduction. Under this framework, a better tradeoff between FAR and FRR was obtained in VAD. Considering the non-stationarity property of speech and noise, we further extended our work to an adaptive regularization framework, i.e., using regularization which controlled the tradeoff between the approximation accuracy and generalization ability of the approximation function. Correspondingly, the tradeoff between noise reduction and speech distortion could be obtained via a regularization parameter. In addition, the regularization was applied in a reproducing kernel Hilbert space (RKHS). It is easy to incorporate a nonlinear transform to explore the nonlinear and high-order statistical structure of speech. It helped to improve the tradeoff for a better VAD. In our previous study, we designed a regularization algorithm with fixed regularization parameter for VAD, and showed promising advantages [6]. Considering the non-stationarity property of speech and noise, we further extend our work to an adaptive regularization framework, i.e., using regularization parameter adaptively to the local variations of signal to noise ratio (SNR) of speech. The VAD designed in this adaptive regularized space will improve the robustness.

We have proposed a regularization framework for noise reduction [6]. In this framework, the speech estimation is regarded as a functional approximation and generalization problem. The solution was obtained via a regularization which controlled the tradeoff between the approximation accuracy and generalization ability of the approximation function. Correspondingly, the tradeoff between noise reduction and speech distortion could be obtained via a regularization parameter. In addition, the regularization was applied in a reproducing kernel Hilbert space (RKHS). It is easy to incorporate a nonlinear transform to explore the nonlinear and high-order statistical structure of speech. It helped to improve the tradeoff for a better VAD. In our previous study, we designed a regularization algorithm with fixed regularization parameter for VAD, and showed promising advantages [6]. Considering the non-stationarity property of speech and noise, we further extend our work to an adaptive regularization framework, i.e., using regularization parameter adaptively to the local variations of signal to noise ratio (SNR) of speech. The VAD designed in this adaptive regularized space will improve the robustness.

The remainder of this paper is organized as follows. Section 2 gives a brief introduction of the regularization framework for signal processing. Section 3 introduces the extended work on adaptive regularization. Section 4 shows the evaluation experiments. Discussions and conclusion are given in section 5.

2. Regularization framework for signal approximation

The estimation of clean speech signal from noisy observation can be regarded as a functional approximation problem. A good

10.21437/Interspeech.2011-678

INTERSPEECH 2011
choice of the approximation function should give good estimation or prediction of clean speech even in adverse noisy conditions. In this section, we introduce the regularization framework for noise reduction, and explain the regularization in a prediction point of view.

2.1. Signal approximation in a reproducing kernel Hilbert space

We suppose that the noisy observation is represented as follows:

\[ y_i = f(x_i) + \epsilon_i, \quad (1) \]

where \( f(x_i) \) is the target function to be approximated (it is suggested that most speech structure is encoded by this function), and \( \epsilon_i \) is noise. From this observation, we try to approximate or learn the target function \( f(\cdot) \) from an observation data set \( S = \{ (x_i, y_i) : i = 1, \ldots, I \} \). \( x_i \in \mathbb{R}^p \) is a vector, and \( y_i \in \mathbb{R} \) is the response or label information (in case of classification tasks). The finding of the function \( f(\cdot) \) is an ill-posed problem as introduced in statistical learning theory. In order to make the problem to be well posed, we suppose that \( f(\cdot) \) is in a reproducing kernel Hilbert space with a certain smoothness that can be used to approximate the speech as follows [9]:

\[ y_i = f(x_i) = w^T \Phi(x_i) \quad (2) \]

where \( \Phi(\cdot) \) is a mapping function that maps a vector to a high dimensional space, and \( w \) is a weighting coefficient vector that uniquely determines the target function \( f(\cdot) \). Hence the problem is to find a mapping function \( f(\cdot) \) by minimizing an objective function \( H(f) \) as follows:

\[ f^* = \arg \min_f H(f) \]

\[ H(f) = \frac{1}{I} \sum_{i=1}^{I} (y_i - f(x_i))^2 + \lambda \| f \|^2_K. \quad (3) \]

There are two components in this objective function \( H(f) \), i.e., the approximation error and smoothness of function \( f(\cdot) \). \( \| f \|^2_K \) is the norm of the function in a RKHS corresponding to a kernel matrix \( K \) constructed from the training data set via a mapping function \( \Phi(\cdot) \). \( \lambda \) is the regularization parameter to control the balance between the approximation error and the smoothness of the function. If \( \lambda \) is small, most of the noise structure is encoded in the approximation function which may cause false acceptance in speech detection. If \( \lambda \) is large, most of the speech structure is removed which may result in false rejection in speech detection.

Based on the representer theorem [9], the solution to Eq. 3 satisfies:

\[ f(x) = \sum_{i=1}^{I} c_i K(x, x_i) \quad (4) \]

In Eq. (4), \( K(\cdot, \cdot) \) is the kernel function which creates a Gram matrix \( K \) with elements defined as follows:

\[ K(x_m, x_m) = \Phi^T(x_m) \Phi(x_m) \quad (5) \]

In real applications, we do not need to know the mapping function \( \Phi(\cdot) \) explicitly. We only need to calculate the inner product of the mapped vectors via a kernel function. The kernel function can be chosen as a Gaussian kernel function, or polynomial function which is widely used in statistical learning field [9]. In Eq. (4), \( c_i \) is the coefficient which depends on the training data samples. By using the representer theorem, the coefficient vector can be obtained by solving the problem in Eq. (3) as follows [9]:

\[ c = (K + \lambda I)^{-1} y, \quad (6) \]

where \( I \) is the identity matrix, the coefficient vector \( c = [c_1, \ldots, c_I]^T \), and the observation vector \( y = [y_1, \ldots, y_I]^T \).

2.2. Approximation function from prediction point of view

In learning an approximation function (shown in Eq. (3)) with an observation sequence \( y_i \), we re-formulate the data in the form of training data pair \((x_i, y_i)\) with the input vector formulated as \( x_i = [y_{i-1}, y_{i-2}, \ldots, y_{i-p}] \), where \( p \) is the dimension of the data vector. For easy understanding, we show the processing in Fig. 1 as a prediction problem. In this figure, the \( \phi_j(\cdot), j = 1, 2, \ldots \) is the basis function. From this figure, we can see that the estimation is a prediction based processing with a nonlinear transform function. The prediction captures the nonlinear and high-order statistical structure of the signal. By implicitly choosing a nonlinear mapping via a kernel function, we can approximate the signal by keeping the nonlinear and high-order statistic information of the signal in a regularized RKHS.

3. Adaptive regularization parameter selection

As shown in Eq. 3, the regularization parameter \( \lambda \) controls the trade-off between the approximation error and generalization ability of the approximation function. In noise reduction, the purpose of the regularization parameter selection is to make the smoothed noisy signal approximate to the clean speech signal (in this study, the estimated speech is obtained from the approximation of \( y \) as \( \hat{y} \)). Many objective measurements can be used to evaluate the goodness of this approximation, for example, in measurement of speech distortion, the Itakura-Saito distance and the log spectral distance (LSD) in dB [7], are often used. In this study, we adopted the LSD between the estimated speech and clean speech as a criterion. The LSD for one frame is defined as follows:

\[ \text{LSD}_\lambda(j) \triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} \left( \frac{1}{10 \log_{10} \left( \hat{Z}^2(j, \omega) / \tilde{Z}^2(j, \omega) \right)} \right)^{\frac{1}{2}} d\omega, \quad (7) \]

where \( j \) is the frame index; \( Z^2(j, \omega) \) and \( \tilde{Z}^2(j, \omega) \) are the power spectrum of clean and estimated speech signals, respectively. They depend on the regularization parameter \( \lambda \). Intuitively, more noise reduction is needed in low SNR conditions,
while less is required in high SNR conditions. Correspondingly, large regularization parameter should be used for reducing more noise, while small regularization parameter should be applied for less noise reduction. Based on this consideration, we design the regularization parameter as a function of local SNR.

A clean speech data set (40 speech utterances) and its noisy ones with SNR 0, 5, 10, and 15 dB are used to estimate the SNR dependent regularization function. In order to exclude the non-speech frames, a silence removing processing is used on clean speech. The calculation was only applied on speech frames for finding the frame based optimal regularization parameter which is defined as follows:

\[
\lambda_{j}^\text{opt} = \arg \min_{\lambda \geq 0} \text{LSD}_{\lambda}(j)
\]  

(8)

Similarly as we did in [6], the polynomial kernel function with nonlinear degree two was used in this optimization (determined through experiments). Based on Eq. 8, the optimal regularization parameter was estimated frame by frame. Finally, we obtained an optimal regularization parameter data set corresponding to local SNR as \( \{ \lambda_{j}^\text{opt}, \text{snr}; j \} = 1, 2, ..., N \}. \) In our study, 16000 frames were used with 32 ms frame length and 16 ms frame rate. The results are shown in Fig. 2. In this figure, the optimal regularization parameter vs the local signal to noise ratio.

![Figure 2: Adaptive regularization parameter vs the local signal to noise ratio.](image)

The optimal regularization parameter and local SNR pairs are marked as “scattered data”. The regularization parameter values are shown with a log transform for better visualization. The average of these scattered data samples is shown as “Empirical curve”. From this figure, we can see that the optimal regularization parameter decreases exponentially with the increase of SNR. Therefore, an exponential function is adopted to fit to these data samples. It is defined as follows:

\[
\lambda_{\text{opt}}^b(\text{snr}) = a \ast \left( c + \exp( b \ast \text{snr}) \right)
\]  

\[= \text{pow}(10, \lambda_{\text{opt}}^b(\text{snr}))\]

(9)

In this definition, \( \lambda_{\text{opt}}^b(\text{snr}) \) is the model fitted regularization parameter in log transformed domain (basis 10 is used for the logarithm), and is transformed back to the original space as \( \lambda_{\text{opt}}(\text{snr}) \) via the power operator ‘pow’ as used in MATLAB. \( a, b, \) and \( c \) are the model parameters, while \( \text{snr} \) is the input local SNR. In this study, by using optimization algorithm, we obtained the parameters as \( a = -1.354, b = 0.8006, c = -0.9094 \), and the input SNR was normalized with mean \(-1.054\) and variance \(10.0\) when using the fitting model for regularization parameter estimation. The model fitted curve is also shown in Fig. 2. Comparing this curve with the “Empirical curve”, we can see that the regularization parameter can be accurately estimated using Eq. (9) when the local SNR is given (in real applications, the local SNR was estimated using the signal subspace based method introduced in [8]).

4. Experiments and evaluations for voice activity detection

The estimated speech signal based on the proposed regularization framework can be regarded as a signal projection in a regularized RKHS. We can design VAD algorithms in this RKHS. For simplicity, we adopt the power energy feature in this RKHS for VAD. The power energy feature is defined as:

\[
E_{\text{RKHS}}(\Delta) = \|f\|_K^2 = w^T w = e^T K e
\]  

(10)

Based on this feature, the Otsu’s method was used for finding an optimal energy threshold level for speech and non-speech classification [10]). The performance of the VAD is expected to be robust in noisy environments since the tradeoff of noise reduction and speech distortion is controlled well in the regularized RKHS.

In VAD experiments, the CENSREC-1-C data corpus was used (a Japanese continuous data corpus designed for testing VAD algorithms in noisy environments [10]). Two data sets, i.e., set A and set B, were used. Set A is composed of four noisy conditions of subway, babble, car and exhibition noise, and set B is composed of another four noisy conditions of restaurant, street, airport, and station noise. In testing, low SNR condition which is composed of noise conditions with SNR 5, 0, and -5 dB were used. In each SNR condition, 104 speech data files were used. Two indexes, false rejection rate (FRR) (rate of the number of speech frames detected as non-speech frames) and false acceptance rate (FAR) (rate of the number of non-speech frames detected as speech frames) were used. By varying the threshold obtained using Otsu’s method [10], we calculated VAD results, and measured the performance in a receiver operating characteristic (ROC) curve based on the FRR and FAR.

![Figure 3: ROC curves for two types of noise conditions with SNR=0 dB.](image)

Fig. 3 shows the ROC curves for two types of noise conditions with SNR=0 dB (subway and street noise from data sets A and B, respectively). In this figure, \( x \)-axis is the value of 100-FAR (%), and \( y \)-axis is the value of 100-FRR (%). The figure legends are: VAD results in original space (ORI) (no noise reduction), in linear signal subspace (LSS) (noise reduction as projection on the estimated linear signal subspace which is a special case in our regularization framework when a linear kernel function is used), in a reproducing kernel Hilbert space with constant regularization (RKHS_C) and adaptive regularization (RKHS_A), respectively. In the RKHS, a polynomial kernel function with nonlinear degree two was used. In all these algorithms, the power energy is used as a feature for VAD. From this figure, we can see that a better tradeoff between FAR and FRR in VAD was obtained in the RKHS than in the traditional linear signal subspace. And the adaptive regularization on the RKHS further improved the performance than that using the constant regularization.

We averaged the VAD results for all noise types (8 types of noise in data sets A and B) for the noisy conditions (with SNR 5, 0, and -5 dB), and show the results in Fig. 4. For comparison, the VAD results using G.729B VAD and ETSI adaptive multi-rate based VAD (denoted as AMR1 and AMR2 for its two options) [11] were also obtained. The results are shown in Fig. 4. From this figure, we can see that the proposed regularization processing with constant regularization (RKHS_C) performed
better than the compared traditional algorithms (G.729B VAD and the ETSI AMR1 and AMR2 VADs), and further improved the VAD performance with adaptive regularization processing (RKHS_A).

5. Conclusion and discussions

In this study, we proposed a regularization framework for VAD in which the estimation of clean speech from noisy observations was regarded as a functional approximation problem. The tradeoff between noise reduction and speech distortion in speech processing was well connected to the tradeoff between functional approximation and generalization. By introducing the smoothness constraint on the mapping function, we could obtain a mapped space in which most of the speech structure is kept while noise structure is smoothed. Because the regularization was done in a RKHS, we could implicitly incorporate a nonlinear transform function in the approximation by introducing a kernel function. Under this framework, the nonlinear and high-order statistic information of speech was incorporated in feature space which resulted in a better tradeoff between false acceptance and false rejection in VAD. Considering the nonstationarity of speech and noise, we further extended the regularization to an adaptive regularization framework. In this extended framework, the noise reduction was done locally according to the local variations of the SNR which could make a better trade-off for the noise reduction and speech distortion. Our experiments confirmed the efficiency of the processing for VAD.

In the proposed algorithm, several issues need further investigation. From theoretical aspect, the VAD performance (false rejection and false acceptance) should be explicitly analyzed with consideration of the selections of the adaptive regularization parameter and the kernel function. From practical aspect, we need to carefully design data corpus for the estimation of the regularization parameters. Also more experiments should be done with comparisons to currently used advanced VAD algorithms. In the future, we will further develop our algorithm by considering all these issues.

6. Acknowledgements

This study is supported by the MASTAR project of the Knowledge Creating Communication Research Center of the National Institute of Information and Communications Technology (NICT), Japan, and a Grant-in-Aid for Young Scientists from the Ministry of Education, Culture, Sports, Science and Technology of Japan (No.22700193).

7. References